Stochastic MM Subspace Algorithms

Matthieu Terris, Emilie Chouzenoux

Stochastic MM Subspace Algorithms

Matthieu Terris*, Emilie Chouzenoux*†
* Center for Visual Computing, CentraleSuplec, 91190 Gif-sur-Yvette, France
† INRIA Saclay, 1 Rue Honor d'Estienne d'Orves, 91120 Palaiseau, France

***Abstract*—In this paper, we propose a version of the MM Subspace algorithm in a stochastic setting. We prove the convergence of the algorithm and show its good practical performances.**

I. INTRODUCTION

Tuning optimally an algorithm is often done w.r.t. a very specific dataset, making it not robust to different problems. Majorize-Minimize (MM) algorithms aim at minimizing a functional F relying on majorant surrogates of F built at each iteration. Such algorithms have good performance in practice, no parameters to tune, and encompass a large number of famous algorithms. The good performance of adaptive stepsize methods in a deterministic setting have encouraged their use in a stochastic setting. Recently, methods such as ADAGRAD or ADAM have shown good performance. Yet, these algorithms lack theoretical guarantees. Our contribution is to propose a version of the MM Subspace algorithm in a stochastic setting.

II. PROPOSED ALGORITHM

In this paper, we consider that we minimize a functional $F : \mathbb{R}^N \to \mathbb{R}$ through an oracle returning noisy observations g_k of the true gradient $\nabla F(x_k)$. A classical MM methods would compute at each iteration a quadratic function

$$h(x, x_k) = g_k^\top (x - x_k) + \frac{1}{2}(x - x_k)^\top A(x_k)(x - x_k) \tag{1}$$

for all $x \in \mathbb{R}^N$ and iterate

$$x_{k+1} = \underset{x \in \mathbb{R}^N}{\operatorname{argmin}}\ h(x, x_k) = x_k - \gamma_k\, A(x_k)^{-1} g_k \tag{2}$$

where γ_k is a decreasing stepsize. The main drawback of these algorithms is the (necessary) inversion of the matrix $A(x_k)$ to solve (2). To circumvent this problem, MM Subspace Algorithms [1] solve the minimization of h within a subspace rather than in the whole space. To do so, one should choose the collection D_k of supporting directions of this subspace. Following [2], we define $D_k = [g_k | x_k - x_{k-1}]$, the subspace version of (2) reads

$$x_{k+1} = x_k - \gamma_k\, D_k (D_k^\top A(x_k) D_k)^{-1} D_k^\top g_k \tag{3}$$

Assuming that the function F is smooth and strongly convex, and under classical assumptions from stochastic approximation, the iterates $(x_k)_{k \in \mathbb{N}}$ generated by the MM Subspace algorithm (3) satisfy

$$\mathbb{E}[F(x_k)] \underset{k \to \infty}{\longrightarrow} 0 \tag{4}$$

and

$$\mathbb{E}[\|x_k - x^*\|^2] \underset{k \to \infty}{\longrightarrow} 0 \tag{5}$$

This result can be extended to the case where F satisfies a global Kurdyka-Łojasiewicz inequality with exponent $\theta \in [1/2, 1]$. A convergence rate in $O(1/k)$ is also obtained with some conditions on the convergence rate of γ_k.

III. APPLICATION TO BINARY CLASSIFICATION

In our stochastic setting, a widely spread application is binary classification. The problem can be formulated as

$$\min_{\theta \in \mathbb{R}^N} F(\theta) = \mathbb{E}[f(x, y, \theta)] \tag{6}$$

where (x, y) is a random variable, whose realizations $(x_k, y_k) \in \mathbb{R}^N \times \{-1; +1\}$ are observations of couples of input-ouputs. In this setting, y contains the labels, and x the features. In this configuration, we choose f as

$$f(x, y, \theta) = \log\left(1 + e^{-y\theta^\top x}\right) + \frac{\mu}{2}\|\theta\|_2^2 \tag{7}$$

Simulations show a good behaviour of the algorithm, in particular when the initialization of the algorithm is done at random. In this case, our algorithm outperforms other commonly used algorithm (cf Figure 1).

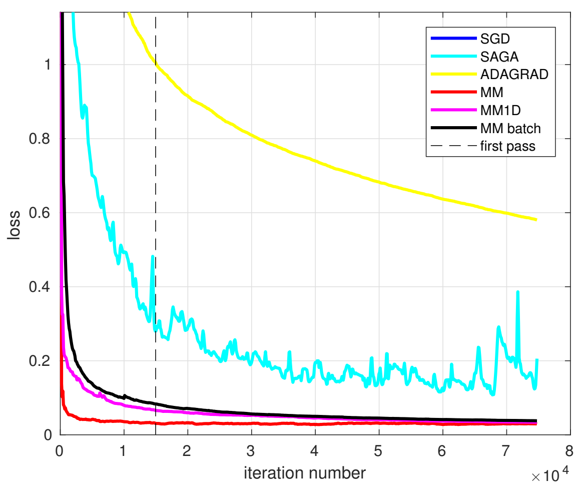

Fig. 1. Convergence of SGD, SAGA, ADAGRAD and different versions of the proposed stochastic MM Subspace algorithm on the W8A dataset [3].

REFERENCES

[1] E. Chouzenoux, J. Idier, and S. Moussaoui, "A majorize–minimize strategy for subspace optimization applied to image restoration," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1517–1528, 2011.

[2] E. Chouzenoux, J.-C. Pesquet, and A. Florescu, "A stochastic 3mg algorithm with application to 2d filter identification," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*. IEEE, 2014, pp. 1587–1591.

[3] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.